Multiclass Cyberbullying Detection Using BERT

Mohammad Yoga Pratama
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
6025241029@student.its.ac.id

Riyanarto Sarno
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Abdullah Faqih Septiyanto
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025221022@student.its.ac.id

Agus Tri Haryono
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7025231020@student.its.ac.id

Shoffi Izza Sabilla
Department of Medical Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
shoffi@its.ac.id

2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C) | 979-8-3315-0857-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/BTS-I2C63534.2024.10941912

***Abstract*— On social media, cyberbullying has grown into a significant issue that impacts people's mental health and well-being. As the language used is often subtle and implicit, it is difficult to detect cyberbullying directly. This study examines the efficacy of transformer-based models, specifically BERT and DistilBERT, to detect cyberbullying compared to conventional machine learning techniques using TF-IDF with Naive Bayes, K-Nearest Neighbors, and Random Forest. To improve the accuracy of the model, we preprocessed the data by normalizing the text, removing stopwords, and expanding slang. Although DistilBERT achieves a competitive 94% accuracy with less computational requirements, the experimental results show that BERT surpasses the standard model with 95% accuracy, 95% precision, 95% recall, and an F1-score of 95%. The results show that transformer models - particularly BERT - are more effective than conventional methods in capturing the nuanced context of coarse language. This research highlights the potential of transformers for accurate cyberbullying detection on social media, with future research aiming to optimize model efficiency and explore ensemble methods to further improve performance.**

Keywords—multi-class, text classification, supervised learning, BERT, DistilBERT

I. INTRODUCTION

Cyberbullying has become a serious problem in the digital era, affecting individuals of all ages across all social media platforms. Cyberbullying is defined as the act of electronic communication to harass, intimidate, or threaten another person, which can have psychological effects on the victim such as anxiety, stress, depression and in the most severe cases, suicide [1] [2]. This anonymous online interaction has led to the rampant action of cyberbullying. Therefore, there is a need for an automated system that is able to detect cyberbullying accurately and quickly, so that it can help reduce the negative impacts of cyberbullying.

Traditional methods for text classification, such as TF-IDF combined with supervised learning algorithms have been widely implemented [3] to handle natural language processing tasks such as sentiment analysis, spam detection, and cyberbullying [4]. Although these methods have proven effective to some extent, they often struggle to capture the more complex semantics and contextual nuances that characterize abusive language in social networks.

In recent years, transformer-based models such as BERT (Bideirectional Encoder Representations from Transformers) have revolutionized NLP by making it possible for machines to comprehend word connections and context in a more sophisticated way. BERT, along with its lighter variant, DistilBERT, leverages deep bidirectional context understanding, allowing for more accurate detection of complex language [5]. These models have shown strong performance on a variety of NLP tasks and are particularly promising for applications requiring high language understanding such as cyberbullying detection. However, the computational overhead of transformer models can be significant, posing challenges in terms of scalability and resource allocation.

This study aims to explore and compare the performance of the Supervised learning model with TF-IDF feature extraction with transformer-based models, namely BERT and DistilBERT to detect cyberbullying. The supervised learning approach combines TF-IDF with the Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest classification models, while the transformer approach uses BERT and DistilBERT to evaluate the performance of deep learning in capturing complex linguistic clues. We evaluate the performance of each model based on accuracy, precision, recall, and f1-score, thus providing insight into the advantages and disadvantages of each model built. Through this comparative study, we hope to demonstrate the potential of the transformer model in improving the accuracy of cyberbullying detection.

II. LITERATURE REVIEW

The increasing cases of cyberbullying in cyberspace and its serious psychological impacts in recent years have made research on cyberbullying detection develop rapidly. One of the main challenges in detecting cyberbullying is understanding the context of messages that often use language explicitly or subtly through sarcasm.

Cyberbullying detection generally uses machine learning techniques with simple feature extraction such as TF-IDF or Bag-of-Words. The study conducted by [6] cyberbullying detection specifically handles comments containing elements of hatred or insults that are not always explicit. The author proposes three machine learning models, namely SVM, NB, and LR. This article mentions the importance of data preprocessing, such as removing stopwords and using POS tangging to determine the class of cyberbullying based on certain keywords such as offensive words, racism, sexual, physical, etc.

Research conducted by [7] introduces CyberBERT fine-tune to detect cyberbullying on social media. The dataset used is from Twitter, Formspring and Wikipedia, this study shows that BERT can classify harmful content such as racism, sexism, and toxic content with better accuracy than other models such as SVM, LR, CNN, and BiLSTM with attention layers. The main advantage of the BERT model in this study is its ability to capture linguistic context bidirectionally, which allows detecting linguistic nuances in cyberbullying texts that are often implicit. This study also uses knowledge distillation to reduce the computational burden without sacrificing accuracy. The results of this study provide comparative knowledge regarding the effectiveness of transformer models in detecting multiclass cyberbullying, where fine-tuning modeling such as BERT is able to improve detection performance compared to supervised learning-based methods such as Naive Bayes and Random Forest with TF-IDF feature extraction.

In article [8] conducted a comprehensive review on deep learning-based cyberbullying detection. This article highlights that deep learning approaches such as Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM), provide better performance than traditional machine learning methods, especially in handling large and unstructured text and image data on social media. This article emphasizes the importance of sophisticated data representations, such as Word2Vec, Glove and BERT that enable deep learning models to better capture linguistic context. This study identifies key challenges in detecting cyberbullying such as limited labeled datasets, difficulties in handling language variations, and the need to consider emotional and cultural elements in the analyzed content.

A. Evaluation Metrics

In this study, we use the main evaluation metrics to assess the performance of each cyberbullying detection model, namely precision, recall, f1-score and accuracy [9] [10]. Each metric has a specific function in measuring the model's ability, the equation can be seen as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy is the value that is correctly classified into its respective class.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Precision is the accuracy of the model in correctly predicting the positive class against all positive predictions made by the model.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall is a measure of how well the model is at identifying all true positive cases.

$$F1 - Score = 2\,x\frac{Precission\ x\ Recall}{Precission + Recall} \quad (4)$$

Finally, f1-score is a combination value of precision and recall.

III. METHODOLOGY

This research uses a dataset of tweets obtained from the Kaggle platform to develop a cyberbullying detection model. This dataset includes various texts that need to be processed and analyzed to recognize potentially cyberbullying content. The first stage in this methodology is data preprocessing, which aims to clean the raw text and improve data consistency. This pre-processing involves several steps: case folding to convert all text to lowercase, removal of excess spaces and irrelevant characters, and removal of stopwords such as “and”, “or”, “which” that have no significance in text analysis. In additional, slang words or abbreviations are replaced with their formal versions, so that the text becomes more consistent and ready for feature extraction.

In Figure 1 after the data is processed, the next stage is feature extraction, where the text is converted into a numerical representation so that it can be utilized by the classification model. In this study, two main approaches are used, namely TF-IDF and transformer-based models such as BERT and DistilBERT. The TF-IDF (Term Frequency-Inverse Document Frequency) approach calculates how often a word appears in a document by considering the uniqueness of the word in the entire dataset [11], and then the results are used as input features for models such as Naive Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Transformers such as BERT and DistilBERT, on the other hand, are used not only for feature extraction, but also as classification models. These two models allow us to obtain a context-rich vector representation, where each word is understood in relation to other words in the sentence, thus providing a more meaningful and accurate representation in cyberbullying detection.

After the features are extracted, the classification process is carried out with the various algorithms mentioned, namely NB, SVM, KNN, as well as the BERT and DistilBERT models. The model aims to identify texts containing cyberbullying elements by classifying the data into the appropriate categories. Evaluation of the model is done by measuring accuracy, precision, recall, and F1-score, as well as using confusion matrix to provide a more detailed picture of the correct and incorrect predictions in each class. Through a systematic series of stages, this methodology is expected to produce an effective and accurate model in detecting cyberbullying in text data.

A. Dataset

The "cyberbullying_tweets" dataset, which was selected by [11] from the Kaggle platform, is used in this work. With an emphasis on cyberbullying, this dataset combines six Twitter datasets (Bretschneider, Chatzakou, Waseem, Davidson, WISC, and Hate Speech). With two columns

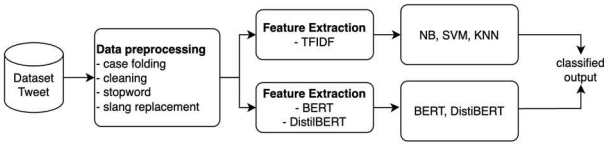

Fig 1. Design Research

TABLE I. SAMPLE OF DATASET

No	tweet_text	cyberbullying_type
1.	U have destroyed your life. Welcome to the world of Islamic Terrorism.	religion
2.	@Foxfairy And I can't think of any good remakes period, actually.	gender
3.	All I need in this life of sin is me and my boyfriend □□	not bullying
4.	why are girls that bullied me in high school trying to add me on facebook??? fuck off	age
5.	I love seeing these racist Hispanics that call black ppl nigger.. Im waiting on the dayyyy I'll educate the fuck out yo dumb ass!	ethnicity

labeled "tweet_text" and "cyberbullying_type," the dataset comprises 47,692 tweets. There are six class groups in this dataset: age, gender, religion, ethnicity, and non-cyberbullying. With over 7000 data points each class, the dataset demonstrates balance. To prevent misunderstanding, the "other cyberbullying" class is left out. The dataset was split into two subsets for model development: a subset test with a ratio of 20% and a train subset with a ratio of 80% [12].

B. Preprocessing

Each tweet comment in the dataset will go through a preprocessing process to clean the data from special characters, URLs, and remove numbers. This aims to reduce noise in the text dataset, and improve model accuracy by focusing on the contextuality of the data.

1. Lower casing, all text in the dataset is converted to lowercase to ensure consistency and reduce unnecessary word variations. This process ensures the model to treat the same word in different case forms as the same word [12].
2. Slang replacement, this is to handle slang or abbreviations commonly used in online conversations, we use a slang dictionary that contains several abbreviated words or phrases such as "*lmao*" (laughing my ass off), "*idk*" (i don't know), and "*btw*" (by the way). Each word in this slang dictionary will be replaced with its full form, so that the model can understand the intended context well [13].
3. Number and Special Character Removal: All numbers and special characters are removed to reduce "noise" in the data. Only letters and spaces are retained to keep the focus on the text content that is relevant to the classification task [14].
4. Stopwords Removal: We use a list of stopwords that contain common words (such as "and", "the", "is") that do not provide significant contextual information in the classification. These words are removed from the text to reduce the complexity of the data and increase the focus on meaningful words [15].
5. Punctuation Removal: To maintain the integrity of the words in the text, all punctuation is removed from the text. This is done to ensure that punctuation does not affect the recognition of patterns in the text by the model [16].

TABLE II. EVALUATION RESULT

Experiment	Metrics			
	Precision	Recall	F1-Score	Accuracy
TFIDF+NB	84%	83%	81%	83%
TFIDF+KNN	78%	65%	67%	65%
TFIDF+RF	94%	93%	93%	93%
BERT+BERT	**95%**	**95%**	**95%**	**95%**
DistilBERT+DistilBERT	94%	94%	94%	94%

IV. EXPERIMENTAL AND RESULT

In this study, we compare the performance of several models for cyberbullying detection using the traditional approach with TF-IDF feature extraction and transformer-based models, namely BERT and DistilBERT. Evaluation is carried out with Accuracy, precison, recall, and F1-Score metrics for each model which can be seen in Table 2.

In the experiment using TF-IDF for the supervised learning model, the best performance was achieved by Random Forest (RF) with a model accuracy of 93%, precision of 94%, recall of 93%, and f1-score of 93%, this model is superior to Naive Bayes and K-Nearest Neighbors with accuracies of 83% and 65% respectively. The use of TFIDF+Naive Baye and K-Nearest Neighbors shows the limitations of the model in capturing more complex contexts from text data.

Transformer-based approaches, namely BERT and DistilBERT, show superior performance compared to traditional models with TF-IDF feature extraction. The BERT classification model achieves Accuracy, recall, precision, and f1-score of 95%. This shows that the BERT model is very effective in understanding bidirectional context and capturing the varying nuances of language in cyberbullying texts.

In Figure 2 confusion matrix of the BERT model, it can be seen that the model has a strong performance in identifying cyberbullying categories with high accuracy, such as the classes "religion", "age", and "ethnicity" have a very high number of correct predictions. This shows the model's ability to understand bullying patterns in these three categories. However, there were several prediction errors, especially in the "not bullying" class, where the BERT model incorrectly identified several texts as forms of cyberbullying. Overall, the BERT model is able to distinguish each class well.

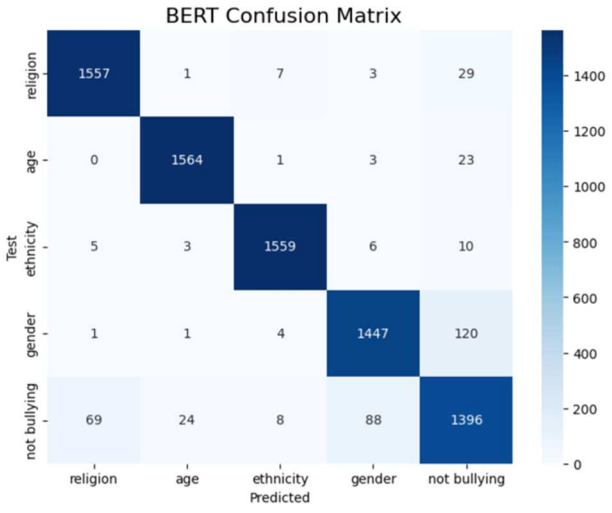

Fig 2. Confusion Matrix BERT

TABLE III. CLASSIFICATION REPORT BERT MODEL

class	accuracy	precision	recall	f1-score
religion	95%	95%	97%	96%
age	95%	98%	98%	98%
ethnicity	95%	99%	98%	99%
gender	95%	94%	92%	93%
not bullying	95%	88%	88%	88%

In Table 3, the BERT model, which is used to detect cyberbullying, performs exceptionally well when it comes to categorizing data into five groups: gender, age, ethnicity, religion, and non-bullying. The classification report states that the model's precision, recall, and f1-score values in the macro average and weighted average were 95% each, and that it obtained an average accuracy of 95%. This demonstrates the model's accuracy and consistency in identifying various forms of cyberbullying. With f1-scores of 98% and 99%, respectively, the age and ethnicity categories outperformed the others, showing that the model is quite good at detecting abusive content based on these factors. The model obtained a f1-score of 96% for the religion category and 93% for the gender category. The BERT model, which is used to detect cyberbullying, performs exceptionally well when it comes to categorizing data into five groups: gender, age, ethnicity, religion, and non-bullying. The classification report states that the model's precision, recall, and f1-score values in the macro average and weighted average were 95% each, and that it obtained an average accuracy of 95%. This demonstrates the model's accuracy and consistency in identifying various forms of cyberbullying. With f1-scores of 98% and 99%, respectively, the age and ethnicity categories outperformed the others, showing that the model is quite good at detecting abusive content based on these factors.

In contrast, the model's f1-score in the non-bullying class was 88%, which was marginally less than that of the other cyberbullying classes. This suggests that while the model still does a good job, it is a little less accurate at identifying data that is not cyberbullying. These findings demonstrate that the BERT model is a powerful tool for category-specific cyberbullying detection because it can accurately identify a wide range of cyberbullying types, particularly in the age and ethnicity categories.

V. CONCLUSION

The impact of cyberbullying is really very serious affecting individuals of all ages across all social media platforms. This anonymous online interaction has led to the rise of cyberbullying. Therefore, there is a need for an automated system that is able to detect cyberbullying accurately and quickly, so that it can help reduce the negative impact of cyberbullying.

The results of this study, which compares transformer-based and TF-IDF models for cyberbullying detection, demonstrate that BERT is superior at categorizing abusive language with 95% accuracy, precision, recall, and F1-score. With 94% accuracy, DistilBERT performs nearly as well but with greater computing efficiency, which makes it a good substitute for large-scale applications. These results support the use of transformer models in effectively detecting various cyberbullying patterns. Future research can focus on optimizing the efficiency of the transformer model or applying ensemble techniques to improve detection accuracy.

ACKNOWLEDGEMENT

This research was funded by Institut Teknologi Sepuluh Nopember (ITS) under Penelitian Kolaborasi Pusat and under the Project Scheme of the Publication Writing and Intellectual Property Rights (IPR) Incentive (Penulisan Publikasi dan Hak Kekayaan Intelektual/PPHKI) Program 2024

REFERENCE

[1] Fattahi, J. Sghaier, F. Mejri, M. Bahroun, S. Ghayoula, R. Manai and Elyes, "Cyberbullying Detection Using Bag-of-Words, TF-IDF, Parallel CNNs and BiLSTM Neural Networks," *IOS Press,* 2024.

[2] E. A. Hendricks and P. T. Tanga, "Effects of bullying on the psychological functioning of victims," *Southern African Journal of Social Work and Social Development,* vol. 31, no. 1, 2019.

[3] Hani, J. Nashaat, M. Ahmed, M. Emad, Z. Amer, E. Mohammed and Ammar, "Social media cyberbullying detection using machine learning," *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 5, pp. 703-707, 2019.

[4] S. T. Rahman, K. H. Mithila and S. Khatun, "An Empirical Study to Detect Cyberbullying with TF-IDF and Machine Learning Algorithms," *Proceedings of International Conference on Electronics, Communications and Information Technology, ICECIT 2021,* no. September, pp. 1-4, 2021.

[5] M. Amgad, A. Ayed, R. M. Gamal and A. Alawi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," *Information (Switzerland),* vol. 14, no. 8, 2023.

[6] Perera, A. Fernando and Pumudu, "Cyberbullying Detection System on Social Media Using Supervised Machine Learning," *Procedia Computer Science,* vol. 239, no. 2023, pp. 506-516, 2024.

[7] P. Sayanta and S. Sriparna, "CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification," *Multimedia Systems,* vol. 28, no. 6, pp. 1897-1904, 2022.

[8] H. M. Tarek, H. M. A. Emran, M. M. S. Hossain, A. Arifa, A. M. Islam and Salekul, "A Review on Deep-Learning-Based Cyberbullying Detection," *Future Internet,* vol. 15, no. 5, pp. 1-47, 2023.

[9] A. Dava, S. Riyanarto, H. S. Chusnul, R. A. Nur and R. Muhammad, "Identification of chronic obstructive pulmonary disease using graph convolutional network in electronic nose," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 34, no. 1, pp. 264-275, 2024.

[10] A. Dava, S. Riyanarto, H. S. Chusnul and R. Muhammad, "Optimization of the Electronic Nose Sensor Array for Asthma Detection Based on Genetic Algorithm," *IEEE Access,* vol. 11, no. July, pp. 74924-74935, 2023.

[11] A. Ravinder, C. Aakarsha, K. Shruti, G. Shaurya and A. Pratyush, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science,* vol. 152, pp. 341-348, 2019.

[12] GeeksForGeeks, "https://www.geeksforgeeks.org," 02 August 2024. [Online]. Available: https://www.geeksforgeeks.org/text-preprocessing-in-python-set-1/. [Accessed 06 November 2024].

[13] A. A. Media and H. M. Abdul, "Preprocessing of Slang Words for Sentiment Analysis on Public Perceptions in Twitter," *IntechOpen,* 2024.

[14] H. A. Tri, S. Riyanarto and A. Rachmad, "Aspect-Based Sentiment Analysis of Financial Headlines and Microblogs Using Semantic Similarity and Bidirectional Long Short-Term Memory," *International Journal of Intelligent Engineering and Systems,* vol. 15, no. 3, pp. 233-241, 2022.

[15] A. Ravinder, C. Aakarsha, K. Shruti, G. Shaurya and A. Pratyush, "The impact of features extraction on the sentiment analysis," *International Conference on Pervasive Computing Advances and Applications –,* vol. 152, pp. 341-348, 2019.

[16] T. MistrY, "ext-PreProcessing — Removing Punctuation and Special Characters," 6 April 2024. [Online]. Available: https://medium.com/@mistrytejasm/text-preprocessing-removing-punctuation-and-special-characters-e3de4cece082. [Accessed 09 November 2024].

[17] Wang, J. Fu, K. Lu and C. Tien, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020,* pp. 1699-1708, 2020.

[18] Salsabila, S. Riyanarto, G. Imam and S. K. Rossa, "Improving cyberbullying detection through multi-level machine learning," *International Journal of Electrical and Computer Engineering,* vol. 14, no. 2, pp. 1779-1787, 2024.

[19] S. R. A. R. Pratama Moch Deny, "Sentiment Analysis User Regarding Hotel Reviews by Aspect Based Using Latent Dirichlet Allocation, Semantic Similarity, and Support Vector Machine Method," *International Journal of Intelligent Engineering and Systems,* vol. 15, no. 3, pp. 514-524, 2022.